Climate Services xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Climate Services

journal homepage: www.elsevier.com/locate/cliser

A strategy to effectively make use of large volumes of climate data for climate change adaptation

Rasmus Benestad *, Kajsa Parding, Andreas Dobler, Abdelkader Mezghani

The Norwegian Meteorological institute, PO Box 43 Blindern, Oslo, Norway

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Big data
Ensembles
Downscaling
Distillation

ABSTRACT

A strategy is suggested for presenting high-resolution temperature maps based on projections from large multi-model ensembles with minimal requirement of data space. This ability to reduce data volumes may be useful for climate services. We present a web-based solution that provides maps with seasonal mean temperatures at 5-min spatial resolution. The maps were generated from downscaled groups of 223 stations from the Barents region, and were based on results from principal component analysis (PCA) for which the five leading modes represented most of the variance and enabled the extraction of salient features while significantly reducing the data volume. A demonstration of the concept showed how different aspects can be distilled, such as ensemble means, ensemble member differences, point-wise time series, probabilities, number of hot/cold days, and various quality aspects. The demonstration included three different types of emissions scenarios: the RCPs 2.6, 4.5, and 8.5. This way of organising data is instrumental to extracting relevant information for decision-making, but does not alone imply actionable adaptation information. The question of reliability and robustness depends on the quality of the data rather than the way it is organised.

Practical Implications

Huge data volumes and different types of data make access, analysis, and distillation (extracting relevant and useful information from the data) is challenging. Data portals have traditionally had a tendency to present climate change projections in terms of a set of multiple single climate model simulations, but it is tricky for users to know which ones to use. Users may select one or a small number of simulations whereas a synthesis derived from a large ensemble may provide more representative information. Handling large climate model ensembles is also computationally demanding. Climate change adaptation and decision-making can benefit from an emphasis on ensemble statistics rather than selected model simulations. Such statistics are more readily obtained through PCA-based strategies which make use of redundancies to reduce the data volume as well as speeding up analytical processes and the estimation of statistics (Benestad et al., 2015). Such techniques also place less emphasis on outlier model results and are designed to highlight the common salient patterns in multi-model ensemble results. These climate models embody a common set of primitive physics-based equations which provide a common ''signal'' in addition to a number of solutions for less well-known aspects such as unresolved processes. The lesser-known processes tend to be solved in various ways and is one source for different model outcomes (''noise''). Here, the PCA may be interpreted as optimising the signal-to-noise ratio, which is assumed to give more reliable results. This is partly supported by the higher skill scores found for multi-model ensemble means (Weigel et al., 2008). This strategy can be used for a wide range of products, such as global climate model (GCM) results, regional climate model (RCM) results, and gridded maps based on empirical-statistical downscaling (ESD). Statistics based on multi-model ensembles can provide the basis for a first guess on probabilities associated with future outcomes on a local scale. These estimates are imperfect since the ensembles are designed in an *ad hoc* fashion regarding the models on which they are based. However, natural and internal variability tend to play a dominant part on a local scale, and large ensembles are able to map their range of outcomes to a reasonable degree (Benestad et al., 2016). The strategy can be seen as step towards distillation in terms of extracting salient information from large data volumes, but does not necessarily imply information that is defensibly robust to the point of action with real world money and real world consequences. Other climate model ensembles and downscaling techniques may potentially produce different information.

* Corresponding author.
E-mail address: rasmus.benestad@met.no (R. Benestad).

http://dx.doi.org/10.1016/j.cliser.2017.06.013
2405-8807/© 2017 Published by Elsevier B.V.

1. Motivation

Climate change adaptation involves risk management for disruptive and dangerous weather events (Kundzewicz et al., 2007) and shifts in the baseline envelope around which long-term infrastructure is designed. One definition of risk R is the product between the consequence of an event $C(x)$ taking place and the probability $P(x)$ for this to happen: $R = C(x)P(x)$ (Intergovernmental Panel on Climate Change, 2014). Climate can be quantified in terms of weather statistics that describe the likelihoods of a range of types of weather situations. For all intents and purposes, climate can be expressed mathematically as the probability distribution function (pdf) for certain single weather variables exceeding a given threshold value $P(X \geqslant x) = 1 - P(X < x)$ (Benestad, 2016). For more advanced multivariate cases, the concept of pdfs can be extended to joint distributions, copulas (Benestad and Haugen, 2007; Schoelzel and Friederichs, 2008), and applied to indices of weather types (Christensen and Bryson, 1966), however, we keep the discussion to the univariate case here for purpose of simplicity. The pdf takes into account both natural variability and systematically forced long-term trends. Its parameters, such as the mean and standard deviation, are expected to change with a climate change [Fig. 2.32, p. 155 (Houghton et al., 2001)], and a warming trend implies a shift in the mean to higher temperature values over time.

Weather fluctuations and natural variability with timescales as long as decades are subject to chaotic non-linear dynamics that render their exact state unpredictable beyond a limited lead time (Lorenz, 1967). Such non-deterministic natural variations are particularly pronounced at regional and local scales which are highly relevant for most types of climate change adaptation (Deser et al., 2012). The non-deterministic nature of internal regional natural variations has been demonstrated with climate models for regional and local spatial scales and on time scales of decades (Deser et al., 2012).

The non-deterministic nature means that it is not possible for a climate model to provide reliable information about the range of possible outcomes based on only one single simulation, regardless of whether the model itself is perfect or not. Although the non-deterministic nature is impossible to predict in a traditional weather forecasting sense, its statistical properties such as the likelihood may be readily predictable. One example to illustrate this is that we can accurately quantify the probabilities of the July temperature in Oslo exceeding the January temperature, even if we cannot predict the exact outcomes. Small ensembles can provide some indication of non-deterministic variability, however, the number of realisations needs to provide a sufficient statistical sample before we can use the results for estimating probabilities. Small samples are subject to "the law of small numbers" with a non-negligible likelihood of a spurious representation of the actual statistics (Kahneman, 2012). In this case, it is the number of independent global climate model (GCM) runs rather than the number of downscaled projections which sets the sample size, as it is the GCM that provides the regional state of natural variability, and downscaled results have a dependency upon the regional conditions. One should keep in mind, however, that the largest uncertainties concerning future climate evolution involve the future greenhouse gas emissions, which depend on how societies and economies evolve and their implications are for the future emissions. In climate modelling, they are described by the representative concentration pathways (RCPs) (IPCC, 2013). The effect of the RCPs can be examined by comparing different GCM simulations following different RCPs to asses how sensitive the future climate is to such differences. Hence, there is a need for a large set of GCM simulations providing a representative sample of the natural variability, model differences, and the effect of different emissions scenarios to capture possible outcomes connected to different phases of internal natural variability.

So-called "climate services" provide knowledge-based facilities for sharing climate know-how with the rest of the society (Vaughan et al., 2016). These climate services need to convey to practitioners that they should use ensembles of climate model simulations which involve a range of different models to capture effects from different model choices. The future climate is to a large extent subject to unknowns connected to different future emissions, and planning for the future needs to use a set of different types of ensembles based on plausible emission scenarios to account for these. One caveat is that multi-model ensembles do not provide a perfect sample, even when based on one emission scenario. They are so-called "ensembles of opportunity" with interdependencies between different models (Sanderson et al., 2015; Smith et al., 2009). Nevertheless, in the absence of other information, and given a dominant contribution of non-deterministic internal variations on the local and regional outcome, such ensembles represent our best starting point for quantifying likelihoods. It is possible to test whether there are systematic differences in the models' ability to represent natural modes of variability, such as the spatio-temporal covariance structure of temperature anomalies. Common empirical orthogonal functions (common EOFs) provide a means to assess how similar these modes are to corresponding ones in reanalyses for the different models (Barnett, 1999).

Some practitioners and decision-makers need to know how often one can expect a winter with mean temperature above or below a specific threshold at a particular location, and empirical-statistical downscaling (ESD) may provide a means to provide a first-estimate for such probabilities through downscaling a large number of global climate simulations such as CMIP5 to a set of locations, subject to the caveat that ensembles do not represent a proper statistical sample (Benestad, 2010). The location of interest is often different to the limited sites where weather observations with long records are located, however, it is possible to produce high-resolution maps from the downscaled data through a gridding procedure (Benestad et al., 2016) (in general, higher resolution may not necessarily imply more information, but in this case, the gridding was carried out using elevation as a co-variate and making use of the relationship between different observations). Hence, large sets of high-resolution maps of seasonal mean temperatures are needed to provide information about the expected number of cold or warm seasons for locations relevant to various practitioners. There may also be needs for different time horizons, seasons, as well as relating the climate information to past changes for validation purposes. This may require maps for seasons spanning long time intervals and calls for a flexible system for climate information delivery (Buontempo et al., 2014).

The utility of data can be enhanced by combining different data sets with tools that can compare and extract common salient features. The combination of different data sets can sometimes yield new information and provide answers to questions that would otherwise been left unanswered. Such combinations are also some of the basis for the "Big Data" concept. One common example is the use of observations for the evaluation of model results, however, the utility is not limited to validation exercises. It is possible to get a first-order estimate of probabilities by analysing the ensemble of projections based on different GCM runs. Regression and Bayesian statistics can be used in more sophisticated analysis, and combining different data may enhance the value of climate services. Such a functionality can both involve data already in the climate data storage (CDS) and can be extended to the practitioners' own data through an upload facility. Any upload facility